

Distance Functions Study in Fuzzy C-Means Core and Reduct Clustering

Joko Eliyanto¹, Sugiyarto Surono²

¹Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia

ARTICLE INFO

Article history:

Received April 04, 2021
 Revised April 19, 2021
 Accepted April 25, 2021

Keywords:

Clustering;
 Core and Reduct;
 Dimensional Reduction;
 Distance Function;
 Fuzzy C-Means

ABSTRACT

Fuzzy clustering aims to produce clusters that take into account the possible membership of each dataset point in a particular cluster. Fuzzy C-Means Clustering Core and Reduct is a fuzzy clustering method is a Fuzzy C-Means Clustering method that has been optimized using the reduction of Core and Reduct dimensions. The method studied is highly dependent on the distance function used. As a further in-depth study, this study was compiled to see the performance of the Fuzzy C-Means Clustering Core and Reduct using various distance functions. We aim to see how consistent the results of this method are across various distance functions and find the best distance function. The seven distance functions are applied to the same dataset. The seven distances are the Euclidean, Manhattan, Minkowski, Chebyshev, Minkowski-Chebyshev, Canberra, and Averages distances. We use UCI Machine Learning datasets for this research. The quality of the clustering results is compared through several measures. Accuracy, Silhouette score, and Davies Bouldin Index are used as internal measurements. The results of Fuzzy C-Means Core and Reduct clustering on all distance functions have significantly decreased computational load. Accuracy and purity values can be maintained with values above 80%. There was an increase in the value of the Silhouette Coefficient Score and a decrease in the Davies Bouldin Index after the application of dimension reduction. This means the quality of the clustering results can be maintained. The distance with the best evaluation result is the Euclidean distance. This method runs consistently across all tested distance functions.

This work is licensed under a [Creative Commons Attribution-Share Alike 4.0](https://creativecommons.org/licenses/by-sa/4.0/)



Sugiyarto Surono,

Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia
 Email: sugiyarto@math.uad.ac.id

1. INTRODUCTION

Current technological developments produce data that is not only large but also continuous. In fact, in recent times, humans have produced more data than all data that has been previously generated [1]. Data at this time is available massively, in large quantities, and in various types [2]. This kind of data is termed big data. This forces us to be able to extract important information from this abundant data.

One of the important pieces of information in the data is the data group. Data grouping is very useful for solving various problems in life. This is often applied as in Customer Segmentation, Recommendation, Image Processing, and others [3]. Often the data clusters have not been previously identified. So, supervised learning cannot be applied. One of the things that can be used as the basis for grouping data is similarity. This method of grouping is called clustering [4]–[6]. The next problem is often data grouped into a group arbitrarily, without considering the possibility to join in other groups. Maybe the computation process will run faster, but its accuracy is questionable. Fuzzy clustering has been proposed as a solution to solve this problem.

The degree of membership is the basis of the fuzzy clustering method [7]. Based on that, say each data point against each exclusion cluster. Fuzzy C-Means clustering is a popular method used in fuzzy clustering [8]. Fuzzy C-Means clustering is a distance-based clustering that applies the concept of fuzzy logic [9]. The clustering process goes hand in hand with the iteration process to minimize the objective function [3][7][8]. The objective function is the sum of the multiplication of the distance between the data points to the nearest

cluster center with the degree of membership [10]. The more iterations, the decreasing the value of the function should be. The distance function used in this method has a key role [11].

Various studies on the effect of distance in the clustering method have been carried out. Some of the results of previous studies that no distance is more dominant and produce outputs that are not much different. The results of clustering are very dependent on the dataset used [3]. Euclidean and Manhattan / City Block, Chebyshev, and Minkowski distances have been identified for their effects in the K-Means Clustering algorithm [11][12]. The results of both studies indicate that the Manhattan distance has a slower computation time than the other distances. In another study, the Euclidean, Manhattan / City Block, Canberra, and Chebyshev distances were applied and evaluated on the fuzzy clustering algorithm [13]–[15]. The results of this study concluded that the results of clustering were very dependent on the data used [16]. In our latest research, the combined Minkowski and Chebyshev distances can also be used to optimize Fuzzy C-Means clustering [17]. Another form of Euclidean distance, namely Average distance, can also be used in the clustering algorithm and produces better results than Euclidean distance [18].

Another way to optimize the clustering method is to apply the dimension reduction method [19]. The dimension reduction method can reduce data dimension but still maintain data characteristics [20]. One of the dimensional reduction methods is Core and Reduct. The Core and Reduct method from the Rough Set theory is proven to be able to improve the performance of Fuzzy C-Means Clustering at the Euclidean distance function [3][21]. In this study, we are doing an expansion of the research on the last results. We want to know whether the consistent application of Core and Reduct can reduce the computational load on Fuzzy C-Means Clustering with various distance functions. The second objective is that we want to find the best distance for the new method. The data used are also limited to five UCI machine learning data, namely iris data, yeast data, seeds data, sonar data, and hill-valley data [22]. In this study, the method is only implemented on numerical data. The Core and Reduct dimension reduction method used was also developed limited to numerical data only.

2. RESEARCH METHOD

2.1. Fuzzy C-Means Core and Reduct Clustering

Fuzzy C-Means Clustering (FCM) is a clustering method that allows certain data can be induced in two or more clusters [7]. This method was invented by Dunn in 1973 and developed further by Bezdek in 1981. The usual application for this method is for pattern introduction. This method is based on minimizing the objective function,

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2, \quad 1 \leq m < \infty \quad (1)$$

where m is a certain real number higher than 1, u_{ij} is the membership degree of x_i , in cluster j , x_i represents the i -th data, c_j represents the cluster centroid j , and $\|\cdot\|$ is the norm which states the similarity between data and the cluster centroid.

Fuzzy partition is applied to trough the continuous optimization process of the objective function which defined before, with the update of the u_{ij} membership degree matrix and c_j cluster centroid by

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (2)$$

and

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m} \quad (3)$$

The iteration will be terminated when $\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon$, where ε is the termination criteria are valued between 0 and 1 and k is the number of the iteration process. This procedure is convergence to minimum local point or saddle point of J_m .

The algorithm of Fuzzy C-Means Clustering consist of these steps [3]:

1. Initialization of $U = \text{matriks}[u_{ij}]$, $U^{(0)}$.

2. In the steps of $-k$, measure the cluster centroid vector of $C^{(k)} = [c_j]$ with $U^{(k)}$ as

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m}$$

3. Update the value of $U^{(k)}, U^{(k+1)}$ as

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. If $\|U^{(k+1)} - U^{(k)}\| < \varepsilon$, then terminate the iteration. If not, back to step 2.

In this method, the value of the objective function and membership degree is much related. During the initial iteration process, we assume that each of the data coordinates already have the value of membership degree for each of the existing cluster. Then, this value will be continuously updated through (2). When (2) is continuously updated, (1) will also keep updated towards its minimum value. One of the challenges of the clustering method is the large computation load. To overcome this, in this method, before entering the clustering process, a dimensional reduction process is carried out using Core and Reduct. The dataset for clustering problems can be viewed as an information table [3]. In the information table, the set of attributes $R \subseteq A$ is called a reduct if R satisfies the following two conditions:

$$1) IND(R) = IND(A); \quad (6)$$

$$2) \forall a \in R, IND(R - \{a\}) \neq IND(A). \quad (7)$$

Condition one states that for each object pair that cannot be distinguished by a subset R, it also cannot be distinguished by A and vice versa. The second condition states that there are object pairs that cannot be distinguished by R - {a} but can be distinguished by A. This means that R is the minimum set of attributes that can maintain the indiscernibility relationship IND (A). Usually, there is more than one reduction in an information table. The set of all reductions from the information table T is denoted as RED (T).

Then, the cores of the attribute set $R \subseteq A$ are as

$$CORE(R) = \bigcap RED(R) \quad (8)$$

Algorithm 1 is Fuzzy C-Means Clustering with Core and Reduct dimensional.

Algorithm 1. Fuzzy C-Means Core & Reduct Clustering

INPUT:

Data input is in the form of variables $X = \{x_1, x_2, L, x_n\}$ and $Y = \{y_1, y_2, L, y_m\}$ expresses objects and attributes. Data is a $n \times m$ matrix, where n is a lot of data and m is the number of data attributes.

PROCESS BEGIN

1. If the dataset is not numeric data, then encoding data is done. If not, then proceed to the next process.
2. Apply the core and reduct method so that the number of variables $Y = \{y_1, y_2, L, y_m\}$ will be a number of new variables $Y = \{y_1, y_2, L, y_p\}$, with $p \leq m$.
3. Applying the fuzzy c-means clustering method so that data clusters are obtained.
4. Cluster evaluation.

PROCESS END

OUTPUT:

The value of the objective function, computational time, purity, Davies Bouldin Index, Silhouette Score and accuracy.

2.2. Distance Function

a. Euclidean Distance

Euclidean distance is known as the most common and applied distance for the Fuzzy C-Means clustering process. For x and y coordinates, this distance is defined as

$$d_{euclidean}(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2} \quad (9)$$

where x_k and y_k are the value of x and y on the certain dimension of n . This distance becomes the standard distance for the fuzzy c-means clustering method [11][18].

b. Manhattan Distance

Manhattan distance is defined as the addition of all of the attributes distance. Hence, for two coordinates data of x and y in dimension n , the Manhattan distance for both of the coordinates is defined as [12]:

$$d_{manhattan}(x, y) = \sum_{k=1}^n |x_k - y_k| \quad (10)$$

where x_k and y_k are the value of x and y on the certain dimension of n .

c. Chebyshev Distance

This distance is also known as the maximum distance, which is defined as the maximum value of the existing attributes distance. The distance for two coordinates data of x and y in dimension n is defined as [12]:

$$d_{chebisev}(x, y) = \max_{k=1}^n |x_k - y_k| \quad (12)$$

where x_k and y_k are the value of x and y on the certain dimension of n .

d. Minkowski Distance

Minkowski distance is the formulation of the metric distance, which defined as [12]:

$$d_{minkowski}(p, x, y) = \sqrt[p]{\sum_{k=1}^n |x_k - y_k|^p} \quad (13)$$

where p is functions as the Minkowski parameter. In Euclidean distance, the value p is equal to 2. For Manhattan distance, the value of p is equal to 1, and for Chebyshev distance, the value of p is equal to ∞ . The metric condition for Minkowski distance is fulfilled as long as $p \geq 1$.

e. Minkowski- Chebyshev Distance

Minkowski-Chebyshev distance is invented by Rodrigues in 2018 [17]. The combination of Minkowski and Chebyshev distance with the weight of w_1 and w_2 is defined as (14). When w_1 is bigger than w_2 , this distance is similar to Minkowski distance, and vice versa. The Minkowski-Chebyshev distance is defined as [23]:

$$\begin{aligned} d_{min-cheb}(w_1, w_2, p, x, y) \\ = w_1 d_{minkowski}(p, x, y) \\ + w_2 d_{chebisev}(x, y) \end{aligned} \quad (14)$$

f. Canberra Distance

Canberra distance is defined as the addition of the absolute value of fraction difference between two data. The formula of this distance is [16]:

$$d_{canberra}(x, y) = \sum_{k=1}^n \frac{|x_k - y_k|}{|x_k| + |y_k|} \quad (15)$$

This distance is very sensitive towards alteration when the value of both of the analyzed coordinates is close to 0. This distance is chosen because of the similarity of its character to Manhattan distance.

g. Average Distance

Average distance is the modification of the Euclidean distance. This modification is applied to improve the clustering result [18]. This distance is defined as:

$$d_{average}(x, y) = \sqrt{\frac{1}{n} \sum_{k=1}^n (x_k - y_k)^2} \quad (16)$$

2.3. Cluster Evaluation

Cluster evaluation is applied to determine the clustering algorithm level of accuracy and the availability of cluster labels. For this research, four tests are applied to evaluate the clustering process, which is purity test, accuracy test, Davies Bouldin Index (DBI), and Silhouette coefficient score.

1. Purity

Purity is used to calculate the purity of a cluster. Purity calculation for each cluster obtained is done by taking the most objects entered in the C-cluster where $1 < i < C$ and C' are the original h-class with $1 < h < C'$. As for the overall purity of the C cluster, it is done by adding up each purity in the C cluster and dividing it by the number of objects defined as follows:

$$\text{purity}(P, C) = \frac{1}{n} \sum_i^c \max_{i \leq h \leq c} |P_i \cap C_h| \quad (17)$$

where $P = \{P_1, P_2, P_3, \dots, P_c\}$ is the cluster set and $C = \{C_1, C_2, C_3, \dots, C_c\}$ is the original class set. Poor clustering has a purity value close to 0. This means that there are no cluster results that match the original class. While a good cluster has a value of purity 1. This means that the cluster results are in accordance with the original class.

2. Accuracy

Accuracy is calculated by adding up the number of objects included in the i -cluster, where $1 < i < C$ the exact class is then divided by the number of data objects. Accuracy is defined as

$$r = \frac{\sum_{i=1}^c a_i}{n} \quad (18)$$

where a_i is the number of objects in the i -cluster that correspond to the original class and n number of n objects. Good accuracy results if all clusters match the original class and then divided by the amount of data will produce a value of 1.

3. Davies Bouldin Index

Davies Bouldin Index (DBI) is one of the methods used to measure cluster validity in a clustering method. The purpose of measurement with DBI is to maximize the distance between clusters (inter-cluster) and to minimize the distance between data points (intra-cluster) in the same cluster [17]. The Davies Bouldin Index (DBI) was introduced by David L. Davies and Donald W. Bouldin in 1979. Before calculating the DBI, first, calculate the variance of each cluster.

$$\text{var}(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

where \bar{x} is the mean of cluster x and n is the number of cluster members.

Then, calculate the Davies-Bouldin Index (DBI) with

$$DBI = \frac{1}{k} \sum_{i=1}^k R_i \quad (19)$$

with $R_i = \max_{j=1, L, k, i \neq j} R_{ij}$ and $R_{ij} = \frac{\text{var}(C_i) + \text{var}(C_j)}{\|c_i - c_j\|}$, where C_i is i -th cluster and c_i is i -th cluster center.

4. Silhouette Coefficient Score

Silhouette Coefficient Score (SCS) is an internal metric that measures the cohesiveness and separation of clusters at the same time [24]. SCS calculates the average distance in a cluster and the minimum distance between an object to another cluster as follows:

$$SCS = \frac{1}{n} \sum_{i=1}^n \frac{\beta_i - \alpha_i}{\max(\alpha_i, \beta_i)} \quad (20)$$

variable α_i is the average distance of objects in a cluster, namely,

$$\alpha_i = \frac{\sum_{j \neq i, x_j \in c_i} |x_i - x_j|}{|c_i|}$$

and β_i represents the distance between an object x_i with the center of the cluster w_j, β_i , which is calculated as

$$\beta_i = \min \{|x_i - w_j|, j = 1, 2, L, k, j \neq i\}.$$

SCS values can range from -1 to 1, ($-1 \leq SCS \leq 1$), where one means the grouping solution is true and -1 means the grouping solution is wrong.

2.4. Dataset

The dataset used in this study is a dataset taken from the UCI Machine Learning website [22]. Table 1 presents a brief description of the dataset used in this study. All of the datasets in Table 1 are numeric types.

Table 1. Dataset

| Dataset | Object | Variables | Class |
|-------------|--------|-----------|-------|
| Iris | 150 | 4 | 3 |
| Seeds | 210 | 7 | 3 |
| Yeast | 1484 | 8 | 10 |
| Sonar | 208 | 60 | 2 |
| Hill Valley | 606 | 100 | 2 |

2.5. Research Method

This research is a numerical simulation using the Fuzzy C-Means Core and Reduct Clustering method. The program is structured using Python 3.0. This clustering method is carried out using seven different distance functions then the results are visualized and analyzed. The research steps are presented in Fig. 1.

3. RESULTS AND DISCUSSION

The initial step of this method is to reduce the dimensions of the dataset. The dataset in Table 1 is reduced using the Core and Reduct method. The result of this dimension reduction is a new dataset with fewer variables. The results of the reduction are presented in Table 2.

Based on the result in Table 2, Core and Reduct work better for the data with high dimensions. As for the low dimension dataset, it tends to be difficult to determine the core of the analyzed data. These results are consistent with research in the same field [21]. Data with low dimensions will make the computation load lower and the computation time can be increased significantly.

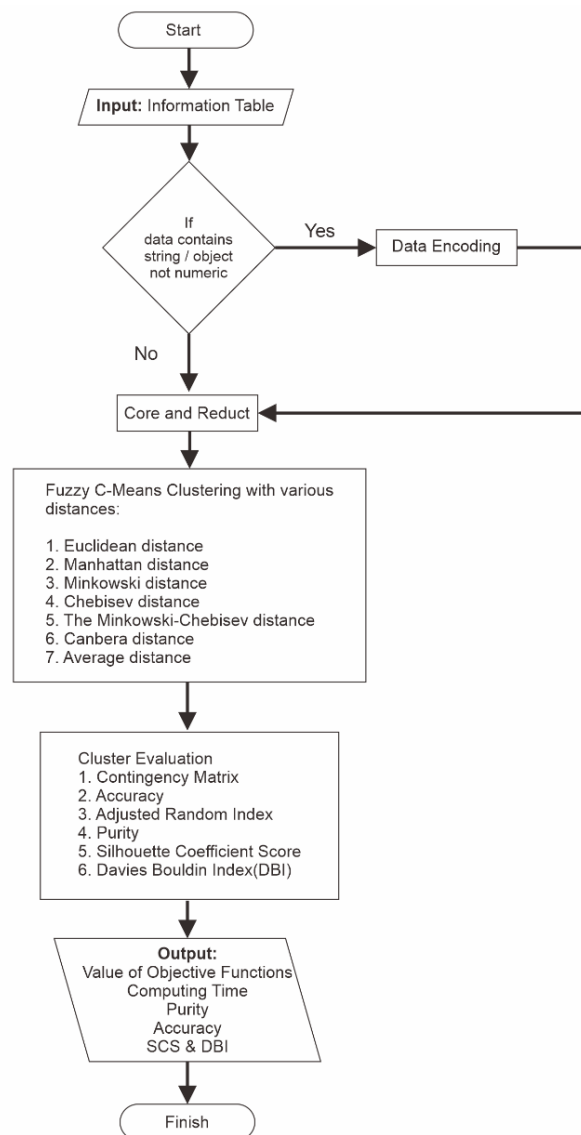


Fig. 1. Research Flowchart

Table 2. Core and Reduct results

| Dataset | Before Reduction | After Reduction |
|-------------|------------------|-----------------|
| Iris | 4 | 3 |
| Seeds | 7 | 2 |
| Yeast | 8 | 4 |
| Sonar | 60 | 2 |
| Hill Valley | 100 | 2 |

The main objective of Fuzzy C-Means clustering is to acquire the objective function value as low as possible. The lower the objective function, the better the result of the fuzzy c-means clustering application. This means that the group of data is more clearly separated. Fig. 2 portrayed the comparison between the values of objective function acquired from Fuzzy-C Means and Fuzzy C-Means Core and Reduct. For five simulated datasets, Core and Reduct able to decrease the value of objective function close to 0% remains for the Euclidean distance, Manhattan distance, Canberra distance, and Minkowski distance. These results support previous research, which states that Euclidean is one of the best performing distances [12].

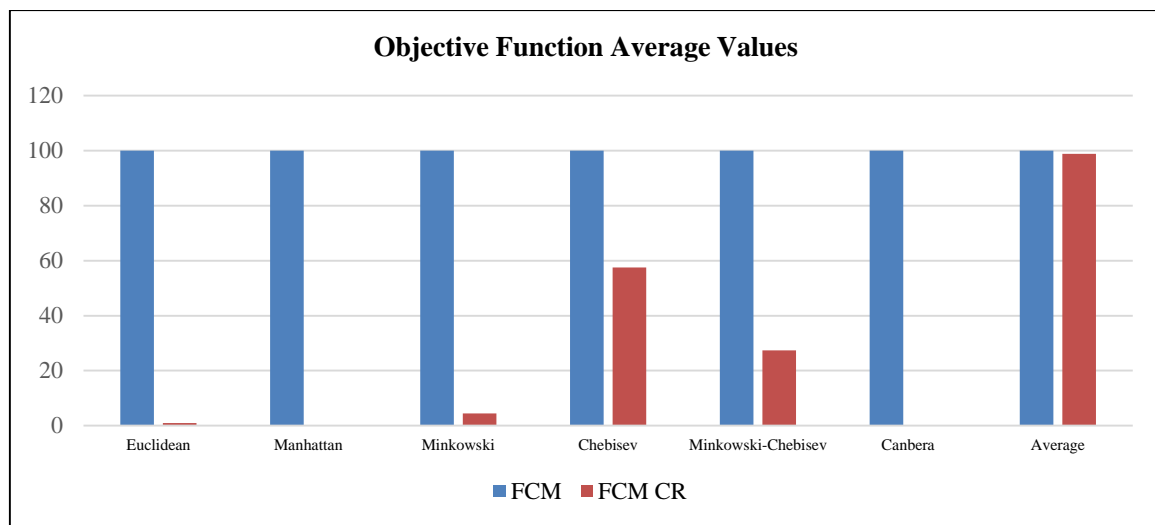


Fig. 2. Objective function average values

The process of computation works in a lower dimension that is resulting in the decrease of the computing weight. This phenomenon will affect the computing time and the number of iteration processes applied. Fig. 3 depicts the comparison of computing time in these two applied Fuzzy C-Means methods. At all distances, the computation time decreased significantly. The highest drop occurred at the Minkowski distance. However, the Euclidean distance still has the lowest computation time of all the distances used. These results indicate the consistency of the Core and Reduct method which is able to reduce the computational load on the Fuzzy C-Means Clustering method. These results complement previous studies which were limited to the Euclidean and Minkowski-Chebyshev distances [3][17].

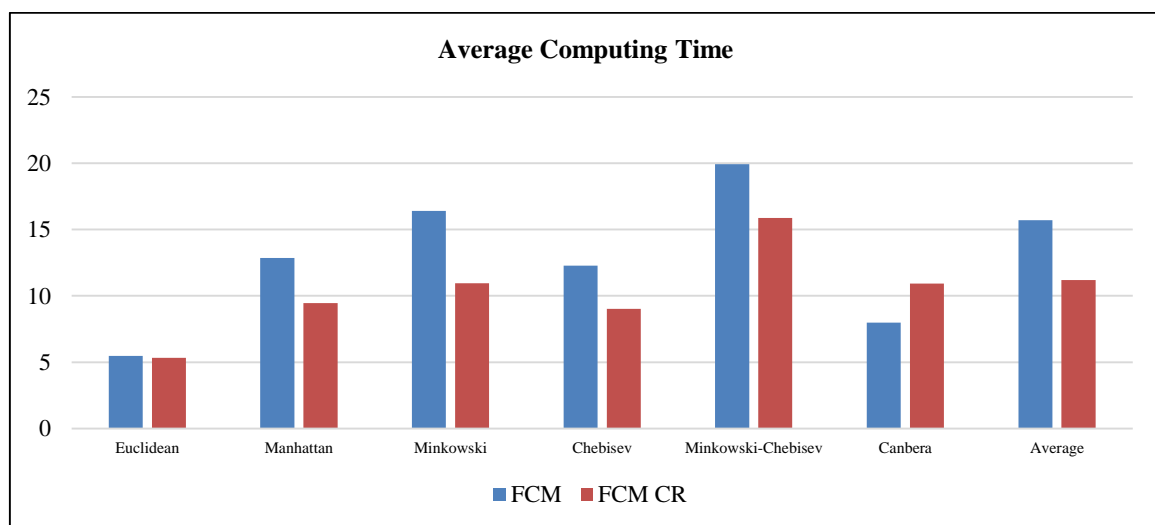


Fig. 3. Average computing time

The lower computation time is usually also due to the iteration time it takes to converge. If the number of iterations is low, then fewer steps must be taken. Fig. 4 illustrates the comparison of the average number of iteration until convergent from both of the analyzed methods. The result shows that the Core and Reduct method is relatively good enough to decrease the computing time of Fuzzy C-Means clustering. This behavior is caused by the decrease of the number of iteration until convergent in the application of Fuzzy C-Means Core and Reduct. The acquired result is not quite significant because Core and Reduct only decrease the number of attributes, while the data record is stagnant.

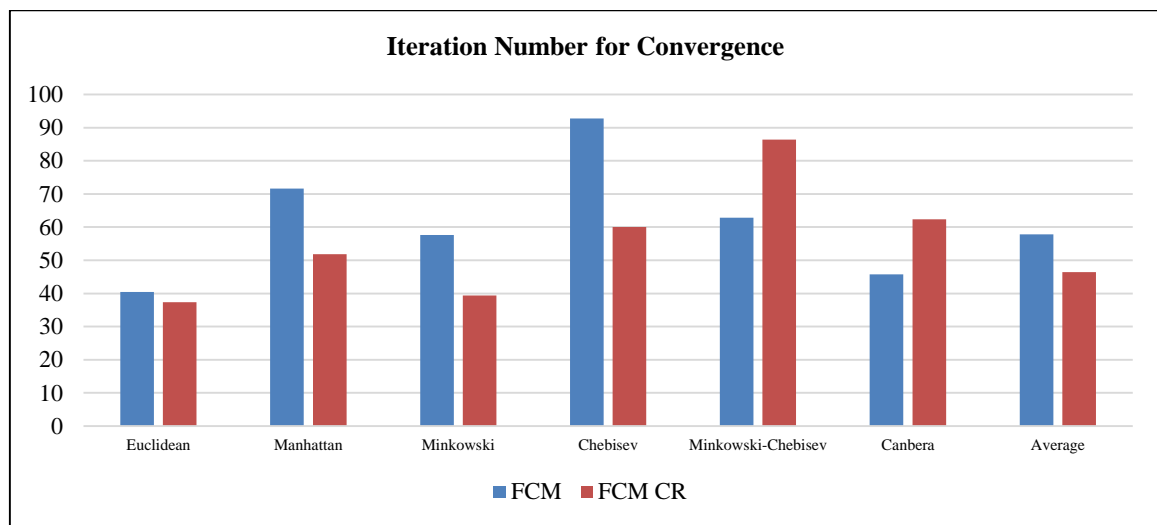


Fig. 4. Number of Iteration Process Until Convergent

Another expected result from the application of dimension reduction is the good quality of clustering result, which has to be similar to the result of the common clustering process. The good cluster output should have accuracy and purity valued close to 1. Accuracy describes the accuracy of the working clustering model. Fig. 5 presents the accuracy value for each distance function. The average accuracy for all distances is 0.47. The distance with the highest accuracy value is Euclidean, with a value of 0.56. Meanwhile, the distance with the lowest accuracy value is the Minkowski distance with a value of 0.38. However, the impact of the application of the reduction of the Core and Reduct dimensions is to compare the accuracy results before and after the dimensional reduction. It can be seen that in Euclidean, Manhattan, Minkowski-Chebyshev, Canberra, the average accuracy can be maintained above 80%. At Minkowski's distance, the accuracy drops drastically until only around 60% remains. This result looks like it can be improved by combining it with Chebysev's distance. This supports the research that Minkowski-Chebyshev's new distance has a great impact and can be used for the optimization of Machine Learning methods [17][23].

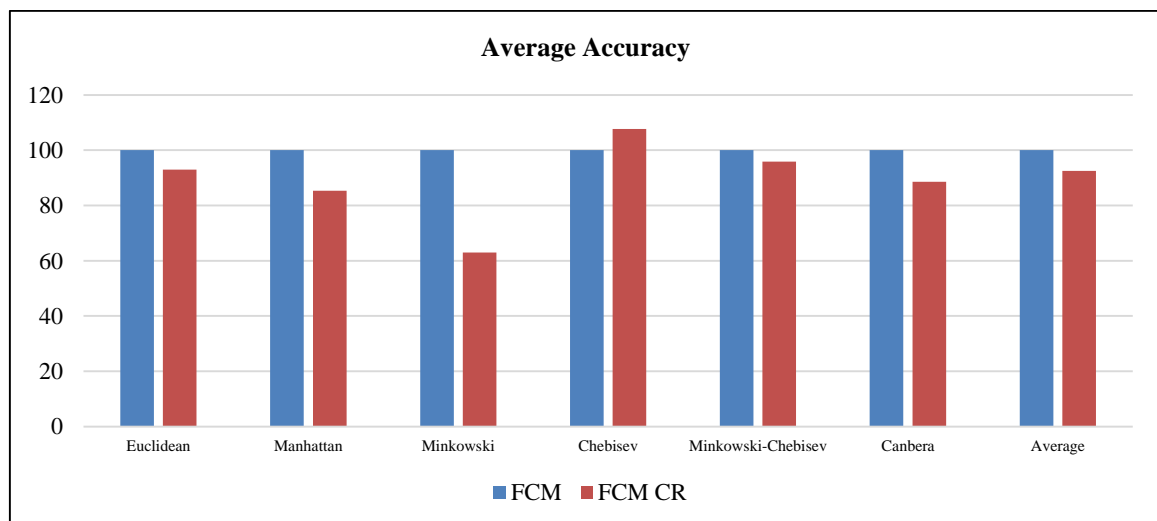


Fig. 5. Average Accuracy

A good cluster result has a purity value close to 1. The result of the average purity value in the Fuzzy C-Means Core and Reduct Clustering method is 0.61. This method succeeded in achieving the highest purity value in the Canberra distance function with an average value of 0.64 and the lowest at the Minkowski-Chebyshev distance with an average value of 0.55. This result is quite good, considering the average purity is more than 0.5. When compared with the Fuzzy C-Means method, the purity value can be maintained very well. The purity value on the reduced data is still above 80% of the original result. This means that even though it is

reduced, the quality of the clusters can be maintained. The result of this paper expands the previous paper [3]. Fig. 6 presents these results.

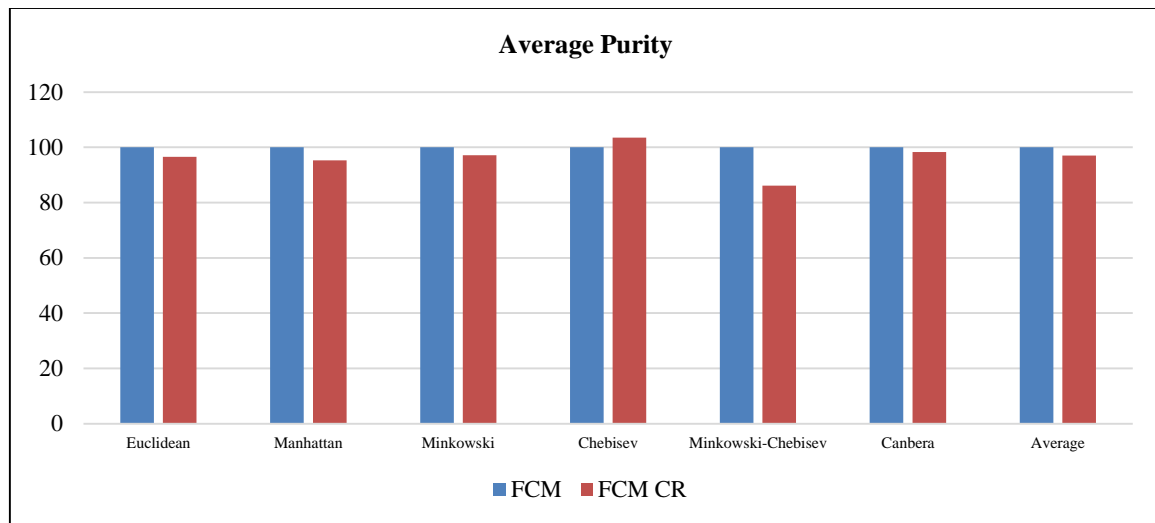


Fig. 6. Average Purity

Another measure of the goodness of a cluster is the Silhouette Coefficient Score. The higher value of this metric, the more correct the clustering results can be. Interesting research results emerge in this section. On average, Fuzzy C-Means Clustering with Core and Reduct application is able to improve the value of silhouette score from 0.399 to 0.507. This means that the Core and Reduct method consistently across all distance functions can improve the quality of cluster results. Fig. 7 presents the Silhouette Coefficient Score value at each distance for the two methods. Nearly all distance functions have a significant increase for this measure. The Minkowski-Chebyshev distance yields the worst score in this case.

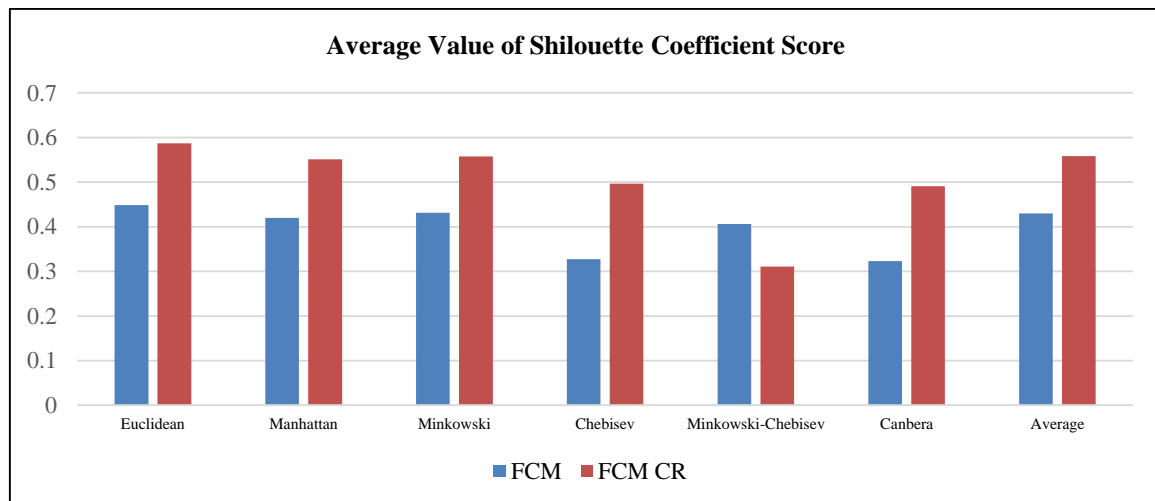


Fig. 7. The average value of the Silhouette Coefficient Score

Fig. 8 shows the comparison of the average DBI results between the Fuzzy C-Means Clustering method and Fuzzy C-Means Core and Reduct Clustering. A cluster will be considered to have an optimal clustering scheme if it has a minimal Davies Bouldin Index (DB) (close to 0) [17]. This new method is able to decrease its remaining value to 53%. This means that the reduction of Core and Reduct dimensions increases the results of Fuzzy C-Means clustering. This applies to all distance functions.

Few of the aforementioned results are linear with the previous research [25][26]. R. Zhao, L. Gu, dan X. Zhu also did research in the same field as this research. Their research resulted in the combination of the C-Means Clustering and Reduct functioned as Rough Set Feature Selection which able to improve the accuracy with a value averaged in 1% [21]. The addition of the Core process applied in this research guaranteed that the result of the dimension reduction was only acquired from the core of the dataset only.

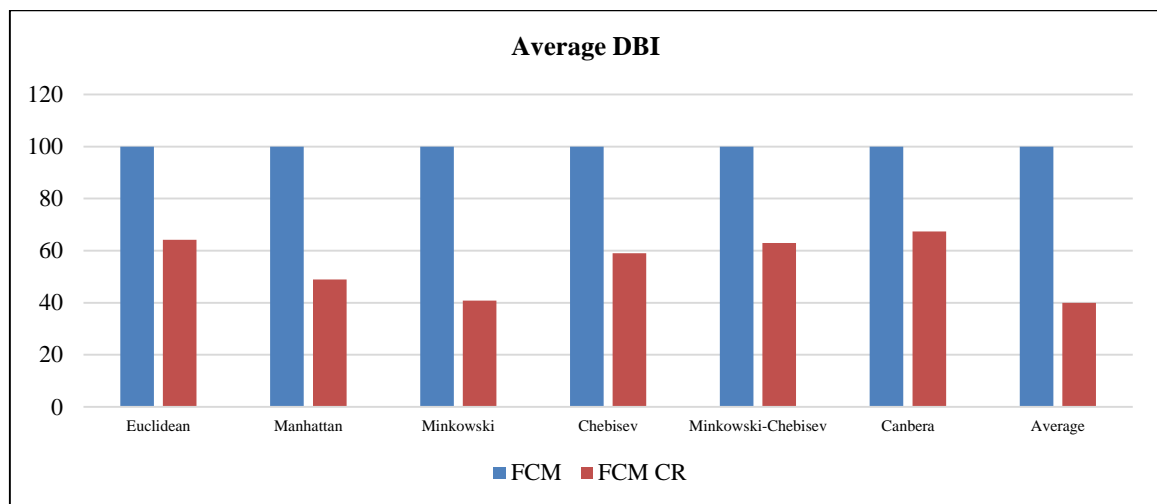


Fig. 8. Average DBI

The best result of this research is acquired on the certain value of U random to initiate the fuzzy c-means algorithm. In the future, an optimization process to determine the initial value for acquiring the best cluster centroid can be developed to further improve the Fuzzy C-Means clustering performance. Table 3 illustrate the valuation of clustering result from the application Fuzzy C-Means Clustering Core and Reduct with seven different distance function parameter. Based on that result, Euclidean distance is still considered the best distance function to be applied.

Table 3. Rank-based Valuation of FCM-CR

| Distance function | Number of Iteration | Computing Time | Objective Function | Accuracy | SCS | DBI | Purity | Total |
|---------------------|---------------------|----------------|--------------------|----------|-----|-----|--------|-------|
| Euclidean | 7 | 7 | 2 | 6 | 7 | 7 | 4 | 45 |
| Manhattan | 2 | 5 | 1 | 7 | 4 | 5 | 5 | 35 |
| Minkowski | 3 | 2 | 3 | 5 | 6 | 2 | 6 | 31 |
| Chebyshev | 1 | 4 | 5 | 1 | 1 | 3 | 1 | 17 |
| Minkowski-Chebyshev | 1 | 1 | 4 | 2 | 3 | 4 | 2 | 19 |
| Canberra | 2 | 6 | 7 | 3 | 2 | 6 | 3 | 32 |
| Average | 1 | 3 | 6 | 4 | 5 | 1 | 7 | 34 |

4. CONCLUSION

The Core and Reduct, dimension reduction method can reduce the computational burden of the Fuzzy C-Means Clustering method on all distance functions. The value of the objective function can be significantly reduced so that the number of iterations and computation time can also be significantly reduced. These results indicate that the reduction of the Core and Reduct dimensions works consistently on the Fuzzy C-Means clustering method with various distance functions. Even so, the quality of the cluster results from this method can still be maintained. These results are shown in the increase in the Silhouette Coefficient Score, the decrease in DBI, and the accuracy and purity values which are still above 80%. Euclidean distance is the best distance with the result of the number of iterations, computation time, the best Silhouette Coefficient Score. The Fuzzy C-Means Clustering method with the reduction of Core and Reduct dimensions is not recommended for the Minkowski-Chebyshev distance function. In the future, research on the development of the Fuzzy C-Means Clustering Core and Reduct method can be applied to image data, video, or other data types.

REFERENCES

- [1] B. Marr, *Big Data in Practice*, 1st ed., vol. 1, no. 1. West Sussex: Wiley, 2016. <https://doi.org/10.1002/9781119278825>
- [2] Y. Riahi and S. Riahi, "Big Data and Big Data Analytics: Concepts, Types and Technologies Big Data and Big Data Analytics: Concepts, Types and Technologies," *Int. J. Res. Eng.*, vol. 5, no. 9, pp. 524–528, 2018. <https://doi.org/10.21276/ijre.2018.5.9.5>
- [3] J. Eliyanto, Sugiyarto, Suparman, I. Djakaria, and M. A. H. Ruhama, "Dimension reduction using core and reduct to

- improve fuzzy C-means clustering performance,” *Technol. Reports Kansai Univ.*, vol. 62, no. 6, pp. 2855–2867, 2020.
- [4] W. Purba, S. Tamba, and J. Saragih, “The effect of mining data k-means clustering toward students profile model drop out potential,” *J. Phys. Conf. Ser.*, vol. 1007, no. 1, pp. 0–6, 2018. <https://doi.org/10.1088/1742-6596/1007/1/012049>
 - [5] D. P. Ismi, S. Panchoo, and Murinto, “K-means clustering based filter feature selection on high dimensional data,” *Int. J. Adv. Intell. Informatics*, vol. 2, no. 1, pp. 38–45, 2016. <https://doi.org/10.26555/ijain.v2i1.54>
 - [6] E. Hardika and S. Atmaja, “Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta,” *Int. J. Appl. Sci. Smart Technol.*, vol. 1, no. 1, pp. 33–44, 2019. <https://doi.org/10.24071/ijasst.v1i1.1859>
 - [7] K. V. Rajkumar, A. Yesubabu, and K. Subrahmanyam, “Fuzzy clustering and Fuzzy C-Means partition cluster analysis and validation studies on a subset of CiteScore dataset,” *Int. J. Electr. Comput. Eng.*, vol. 9, no. 4, pp. 2760–2770, 2019. <https://doi.org/10.11591/ijece.v9i4.pp2760-2770>
 - [8] A. Gosain and S. Dahiya, “Performance Analysis of Various Fuzzy Clustering Algorithms: A Review,” *Procedia Comput. Sci.*, vol. 79, pp. 100–111, 2016. <https://doi.org/10.1016/j.procs.2016.03.014>
 - [9] A. A. Hussain Hassan, W. M. Shah, M. F. I. Othman, and H. A. H. Hassan, “Evaluate the performance of K-Means and the fuzzy C-Means algorithms to formation balanced clusters in wireless sensor networks,” *Int. J. Electr. Comput. Eng.*, vol. 10, no. 2, pp. 1515–1523, 2020. <https://doi.org/10.11591/ijece.v10i2.pp1515-1523>
 - [10] A. Nurzahputra, M. A. Muslim, and R. Kurniawan, “Online Fuzzy C-Means Clustering for Lecturer Performance Assessment Based on National and International Journal Publication,” in *International Conference on Mathematics, Science, and Education*, 2016.
 - [11] S. Kapil and M. Chawla, “Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics,” *1st IEEE Int. Conf. Power Electron. Intell. Control Energy Syst.*, pp. 1–4, 2016. <https://doi.org/10.5120/19360-0929>
 - [12] M. K. Arzoo, A. Prof, and K. Rathod, “K-Means algorithm with different distance metrics in spatial data mining with uses of NetBeans IDE 8.2,” *Int. Res. J. Eng. Technol.*, vol. 4, no. 4, pp. 2363–2368, 2017.
 - [13] B. Charulatha, P. Rodrigues, T. Chitralekha, and A. Rajaraman, “A Comparative study of different distance metrics that can be used in Fuzzy Clustering Algorithms,” *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 2013, pp. 2–5, 2013.
 - [14] A. Singh, A. Yadav, and A. Rana, “K-means with Three Different Distance Metrics,” *Int. J. Comput. Appl.*, vol. 67, no. 10, pp. 13–17, 2013. <https://doi.org/10.5120/11430-6785>
 - [15] P. Grabusts, “The choice of metrics for clustering algorithms,” *Vide. Tehnol. Resur. - Environ. Technol. Resour.*, vol. 2, no. 1, pp. 70–78, 2011. <https://doi.org/10.17770/etr2011vol2.973>
 - [16] Mahatme and Boyar, “Impact of Distance Metrics on the Performace of K-Means and Fuzzy C-means Clustering - an Approach to Assess Student’s performance in E-Learning Environment,” *International Journal of Advanced Research in Computer Science*, vol. 9, no. 1, pp. 888–892, 2018. <https://doi.org/10.26483/ijarcs.v9i1.5417>
 - [17] S. Surono and R. D. A. Putri, “Optimization of Fuzzy C-Means Clustering Algorithm with Combination of Minkowski and Chebyshev Distance Using Principal Component Analysis,” *Int. J. Fuzzy Syst.*, vol. 23, no. 1, pp. 139–144, 2020. <https://doi.org/10.1007/s40815-020-00997-5>
 - [18] A. S. Shirkorshidi, S. Aghabozorgi, and T. Ying Wah, “A Comparison study on similarity and dissimilarity measures in clustering continuous data,” *PLoS One*, vol. 10, no. 12, pp. 1–20, 2015. <https://doi.org/10.1371/journal.pone.0144059>
 - [19] B. R. A. Moreira *et al.*, “Classifying Hybrids of Energy Cane for Production of Bioethanol and Cogeneration of Biomass-Based Electricity by Principal Component Analysis-Linked Fuzzy C-Means Clustering Algorithm,” *J. Agric. Sci.*, vol. 11, no. 14, p. 246, 2019. <https://doi.org/10.5539/jas.v11n14p246>
 - [20] M. M. Deris, N. Senan, Z. Abdullah, R. Mamat, and B. Handaga, “Dimensional reduction using conditional entropy for incomplete information systems,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11657, pp. 263–272, 2019. https://doi.org/10.1007/978-3-030-25636-4_21
 - [21] R. Zhao, L. Gu, and X. Zhu, “Combining fuzzy C-means clustering with fuzzy rough feature selection,” *Appl. Sci.*, vol. 9, no. 4, 2019. <https://doi.org/10.3390/app9040679>
 - [22] D. Dua and C. Graff, “UCI Machine Learning Repository.” University of California, School of Information and Computer Science., Irvine, 2019. <http://archive.ics.uci.edu/ml>
 - [23] E. O. Rodrigues, “Combining Minkowski and Chebyshev: New Distances Proposal and Survey of Distances Metrics Using K-Nearest Neighbours Classifier,” *Elsevier*, 2018. <https://doi.org/10.1016/j.patrec.2018.03.021>
 - [24] F. Wang, H. H. Franco-Penya, J. D. Kelleher, J. Pugh, and R. Ross, “An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10358, 2017. https://doi.org/10.1007/978-3-319-62416-7_21
 - [25] M. F. Dzulkalnine and R. Sallehuddin, “Missing data imputation with fuzzy feature selection for diabetes dataset,” *SN Appl. Sci.*, vol. 1, no. 4, 2019. <https://doi.org/10.1007/s42452-019-0383-x>
 - [26] M. Sammany and T. Medhat, “Dimensionality reduction using rough set approach for two neural networks-based applications,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4585 LNAI, pp. 639–647, 2007. https://doi.org/10.1007/978-3-540-73451-2_67

BIOGRAPHY OF AUTHORS

Joko Eliyanto is an Instructor & Researcher at the Ahmad Dahlan University Center for Data Science Studies. He received his bachelor's degree in Mathematics and master's degree in the Mathematics Education from Ahmad Dahlan University. He is interested in Data Science, Machine Learning, and Artificial Intelligence. He gets TensorFlow Developer Certification since 2020.

Sugiyarto Surono completed his undergraduate education in Mathematics at Gadjah Mada University and completed his master's degree in Mathematics and Statistics at the same university. He has completed his doctorate degree at Universiti Teknologi Malaysia, majoring in optimization. Currently he is working as a lecturer at the Faculty of Applied Science and Technology at Ahmad Dahlan University while doing some research in fuzzy learning for machine learning and deep learning.